Abstract of the Disclosur

A method and system for dynamically allocating cache space in a storage system among multiple workload classes each having a unique set of quality-of-service (QoS) requirements. The invention dynamically adapts the space allocated to each class depending upon the observed response time for each class and the observed temporal locality in each class. The dynamic allocation is achieved by maintaining a history of recently evicted pages for each class, determining a future cache size for the class based on the history and the QoS requirements where the future cache size might be different than a current cache size for the class, determining whether the QoS requirements for the class are being met, and adjusting the future cache size to maximize the number of classes in which the QoS requirements are met. The future cache sizes are increased for the classes whose QoS requirements are not met while they are decreased for those whose QoS requirements are met.